Automatic Assessment of Information Disclosure Quality in Chinese Annual Reports

Xin Ying Qiu[1], Shengyi Jiang[1], and Kebin Deng[2]

[1] CISCO School of Informatics
Guangdong University of Foreign Studies, Guangzhou, China
[2] School of Finance
Guangdong University of Foreign Studies, Guangzhou, China

Abstract. Information disclosure in annual reports is a mandatory requirement for publicly traded companies in China. The quality of information disclosure will reduce information asymmetry and therefore support market efficiency. Currently, the evaluation of the information disclosure quality in Chinese reports is conducted manually. It remains an untapped field for NLP and text mining community. The goal of this paper is to develop automatic assessment system for information disclosure quality in Chinese annual reports. Our assessment system framework incorporates different technologies including Chinese document modeling, Chinese readability index construction, and multi-class classification. Our explorative and systematic experiment results show that: 1) our automatic assessment system can produce solid predictive accuracy for disclosure quality, especially in "excellent" and "fail" categories; 2) our system for Chinese annual reports assessment achieves better predictive accuracy in certain perspective than the counterparts of the English annual reports prediction; 3) our readability index for Chinese documents, as well as other findings from system performance, may provide enlightenment for a better understanding about the quality features of Chinese company annual reports.

Keywords: Text classification, Natural language processing, Information disclosure quality, Application.

1 Introduction

Publicly traded companies in China are required to disclose important information about their companies annually to its investors on the market. Mandatory information disclosure includes company's financial performance, changes in strategies, explanations for such changes, and projections for future performance. The clarity and completeness in information disclosure have an important impact on reducing information asymmetry and therefore improving market efficiency. Historically, Chinese annual reports are mainly studied by researchers in finance, economics, and accounting fields in China. Their study have been focusing on how the quality of annual reports have caused economic consequences or impact corporate governance [1,2]. The quality of information disclosure has

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 288–298, 2013.

been deemed as an important factor in both economic practices as well as academic research. However, the assessment of the quality of information disclosure remains a manual and time consuming process in China. Analysts manually evaluate the quality of Chinese reports each year to assign a grade category for each report. The study of the quality of information disclosure, especially pertaining to Chinese reports, remains an untapped area to the text mining and natural language processing communities.

The application of computer science research in the area of disclosure quality was first proposed by Core (2001) [3]. He suggested that computing the measure of disclosure quality could greatly benefit from the techniques of other research areas such as computer science, computational linguistics, and artificial intelligence. Some relevant works in this direction are those of Davis, Piger, and Seor (2006)[4], Li (2008, 2010)[5,6], Kogan, Levin, Routledge, Sagi, and Smith (2009)[7], Feldman, Govindaraj, Livnat, and Segal (2010)[8], and Lehavy, Li, and Merkley (2011)[9]. Davis et al.(2006)[4] showed that the positive or negative tone in earnings press releases is associated with firms future performance, and captured in market returns. Kogan et al.[7] apply regression techniques to annual reports to construct models for the financial risk level for the period following the reports. Their model results outperform past volatility and are more accurate for annual reports after the Sarbanes-Oxley Act. In F. Li (2010)[6], naive Bayesian machine learning algorithm was applied to study how the information contained in the forward-looking statements in annual reports are related to different financial indicators. Feldman et al. (2010)[8] used regression analysis to show that tone changes in annual reports are associated with immediate market reactions and can be used to predict future stock prices. In general, these studies have focused on specific features of company reports, such as readability, positive and negative tone, and risk level, in stead of the overall quality assessment and its impact.

The SEC (Securities and Exchange Commission) used to conduct manual quality assessment by analysts for annual reports in the US. Researchers in accounting and finance domains have explored this data to study how the quality of mandatory disclosure is related to the forecast of company performance in the US. For example, Gelb and Zarowin [10] empirically confirmed that high disclosure firms provided greater stock price informativeness to the investors. However, these studies relied on the ratings from analysts' manual evaluation, which are no longer available after 1996. Otherwise, such quality index study relies on data of a smaller sample size from labor-intensive document analysis process. In China, analysts' evaluation of annual reports disclosure quality are available for all companies traded at the Shenzhen Stock Exchange. Researchers in the accounting and finance field have explored the disclosure quality ratings to study how disclosure quality is related to cost of equity capital (Wang and Jiang 2004 [1]), corporate governance (Wang and Shen) [11], and stock liquidity (Chen 2007 [12]). These studies mainly focus on the association of quality measure with other economic, managerial, or financial indicators. The methods these studies employ are generally semi-automatic, including content analysis,

manual annotation and categorization, linear discriminant analysis, logit model and other statistical analysis.

We observe from the above literature analysis that automatic assessment of the disclosure quality in Chinese annual reports remains an open research question untapped by the text mining and NLP community. Our overall research goal is to explore the feasibility of applying text categorization methods in constructing automatic models for evaluating Chinese annual reports quality. We believe the significance of such study is three-fold: 1) the development of automatic methods for disclosure quality assessment can supplement the expensive and labor intensive manual evaluation process currently in place; 2) the assessment system can discover the important language and document-level features related to disclosure quality, instead of predefining ex ante limited textual features for further analysis; 3) our results could be compared with those of the more mature study of English annual reports to shed lights on the better understanding of how disclosure quality may be perceived and utilized in different country and economy.

We propose to address our research goals with the following approaches: 1) We use a multi-class text categorization approach and the quality rating data from Shenzhen Stock Exchange to build quality assessment model. Model performance is evaluated with accuracy, and analyzed according to different term weighting schemes, and per-class evaluation. Performance is further compared with the relevant counterpart of English annual reports. 2) Since annual report readability is one of the most popular features in English annual reports[5,9], we implement a Chinese document readability index and evaluate the association between readability measurement and analysts effort. Overall, our paper contributes a foundation of both methodology and results on automatic assessment and analysis of Chinese annual reports quality. The rest of the paper is organized as follows. First, we present our methodologies and experiment design. Next, we analyze our results addressing from our approaches. Our conclusions and directions for future research are then presented at the end.

2 Methodology and Design

Our hypothesis is that we could construct automatic system to assess Chinese annual reports' quality, as a supplement to analysts' manual evaluation. We formulate our design to build automatic assessment system with a multi-class classifier approach. We use the analysts manual quality ratings for annual reports at Shenzhen Stock Exchange as our gold standard. To validate the system's feasibility and evaluate the model's performance, we conduct a series of stratified cross-validation experiments. The details of this approach is presented as follows. We pick readability as a special feature to consider as it has been studied in depth in English annual reports analysis [5,9]. We implemented a Chinese readability index and report results from a regression model to evaluate its association with analysts effort. Our study of how readability and its component features are associated with disclosure quality is currently under way.

2.1 Data Collection and Class Definitions

We automatically retrieved all the Chinese annual reports with disclosure quality ratings for companies traded at the Shenzhen Stock Exchange from 2001 to 2009. After filtering out reports with errors, we obtain a sample set of a total of 4753 company annual reports with manual quality rating data spanning from 2001 to 2009. The distribution of the reports along with quality ratings is indicated in Table 1.

Table 1. Distribution of Annual Reports with Quality Assessment

Year	Number of Docs	Excellent	Good	Pass	Fail
2001	420	28	169	198	25
2002	434	32	204	166	32
2003	461	39	245	155	22
2004	452	28	281	126	17
2005	332	25	176	106	25
2006	538	53	289	170	26
2007	637	62	336	215	24
2008	715	77	432	191	15
2009	764	93	521	134	16
Total	4753	437	2653	1461	202

2.2 Readability Index

Readability is one of the interesting index in the study of English annual reports. Researchers have found out that reports with firms with low readability (i.e. hard to read) have lower earnings [5], and higher number of analysts following [9]. Our goal is to discover how Chinese report readability is associated with disclosure quality, and whether the association between Chinese reports readability and analysts efforts is the same as with English reports. We adopt a readability index as proposed by Yang [13] which has been applied to Chinese documents in other studies. We use the 7-factor and the 3-factor calculations as follow:

$$\begin{aligned} 7-factor\ readability: \quad Y = {} & 13.90963 + 1.54461 \times FULLSEN + \\ & 39.01497 \times WORDLIST - 2.52206 \times STROKES - \\ & 0.29809 \times COUNT5 + 0.36192 \times COUNT12 + \\ & 0.99363 \times COUNT22 - 1.64671 \times COUNT25 \end{aligned} \tag{1}$$

$$\begin{aligned} 3-factor\ readability: \quad Y = {} & 14.95961 + 39.07746 \times WORDLIST + \\ & 1.11506 \times FULLSEN - 2.48491 \times STROKES \end{aligned} \tag{2}$$

where $STROKES$ is the average number of strokes of the Chinese characters in each document; $WORDLIST$ is the proportion of words in the basic word list for each document; $FULLSEN$ is the proportion of full sentences in all sentences in each document; $COUNT5$ is the proportion of 5 strokes characters in all characters for each document; $COUNT12$, $COUNT22$ and $COUNT23$ are calculated similarly as $COUNT5$ but for 12 strokes, 22 strokes and 23 strokes characters respectively.

As a note to our calculation of $WORDLIST$ factor, in absence of a "basic word list" from the original readability paper by Yang, we construct our own basic word list using the vocabulary lists of HSK (Hanyu Shuiping Kaoshi). There are 4 levels of vocabulary for HSK. We use the first three levels to construct a basic word list of 5081 terms (including both single word and multi-word Chinese terms), the size of which is closest to that of the original basic word list.

2.3 Document Models

In information retrieval, documents are typically modeled as vectors of terms with weighting for each term to indicate the importance of term in contributing to the documents' main content. Our research goal is to explore potential indicating textual features that may characterize the different qualities in Chinese annual reports. Besides analyzing certain popular disclosure features such as readability as in Section 2.2, we intend for this baseline system an approach to adopt the typical Bag-Of-Word representation model with TF*IDF weighting scheme. This model is the most successful and widely used where the positions of terms are ignored and the term weighting scheme measures the descriptive information contained in terms.

In Chinese language, terms may compose of single words as well as multi-word phrases. In our pilot study, we experimented with two approaches. One is to use Lucene system and Lucene's ICTCLAS dictionary to segment and index documents. Second is to first use ICTCLAS tool to first segment the documents and then index them with Lucene. We do not observe significant difference in the indexing results and therefore adopt the first approach using Lucene alone. Our indexing experiment originally extracted 54701 terms (including single word and multi-word terms). We observe that some features extracted were meaningless symbols and alphabet combinations. We did a coarse automatic filtering to preserve a feature set of 37809 Chinese terms.

For the TF*IDF weighting schemes, we experimented with 4 variations, namely "atn", "atc", "ltn", and "ltc". The "ltn" and "atn" weights are calculated as follows:

$$ltn: \quad w_i = (\ln(tf) + 1.0) \times \ln(\frac{N}{n}) \tag{3}$$

$$atn: \quad w_i = (0.5 + 0.5 \times \frac{tf}{maxtf}) \times \ln(\frac{N}{n}) \tag{4}$$

where tf is raw term frequency; $maxtf$ the highest term frequency in the document; N is the total number of documents in the collection; n is the number of

documents containing term i; w_i is the weight of term i. The difference between "atc" and "atn", and between "ltc" and ltn" weights are in the normalization factor only such that $weight(term_i) = \frac{w_i}{\sqrt{\sum_i w_i^2}}$, where w_i is either the "ltn" or the "atn" weight as stated above.

2.4 Classifier, Regression, and Experiment Design

Our quality assessment model is based on SVM classifiers. Since we have a four-class categorization problem, we need to consider different options. First, we could perform a one-against-rest classification for each class and combine the results to make a final decision. Second, we could perform a one-against-one classification for n(n - 1)/2 pairs of classes, and combine the results to make a final decision. Third, we could use algorithms designed specifically for multi-class classification. Currently, this article reports results for the first option, as the experiments for options two and three are under way.

For Option one, we use linear SVM to produce three one against-rest models. There are two variants of this in terms of how we combine the results of the three models. First, since we use three binary classifiers to predict the three classes of outperforming, average, and underperforming, each firm will have three scores assigned to it by each of the three classifiers. Our first strategy for combining is to use the highest score to assign a class label. We denote this model as SVM-score. Second, we use LinPlatts method (Platt, 1999)[14] to transform each of the three scores into a probability that the firm belongs to one of the three classes. Then, we use the highest probability to assign a class label to the firm. We denote this model as SVM-prob. We split all the 4753 documents into 10 sets with stratification, so that each the class distribution of each set is equivalent. We perform 10-fold cross validation with these 10 sets of data. Average accuracies are computed for all folds as well as for each binary classification for each of the four classes.

Our regression analysis of how Chinese readability is associated with analysts effort emanates from the study by Lehavy [9] on the English reports. Our hypothesis is that disclosure readability is positively related to number of analysts following, as in the following model:

$$\begin{aligned} Analysts = \beta_0 + \beta_1 Readability_{i,t} + \beta_2 Logsize_{i,t-1} + \beta_4 Lsegments_{i,t} \\ + \beta_5 Std_red_{i,t} + \beta_6 Growth_{i,t} + \beta_7 ADV_{i,t} + \beta_8 Mfcount_{i,t} \\ + \eta_i + g_t + v_{i,t} \end{aligned} \tag{5}$$

where $Analysts$ is the number of analysts following a firm; $Logsize_{i,t-1}$ is the size of a firm; $Lsegments_{i,t}$ is number of reported business segments prior fiscal year; $Std_red_{i,t}$ is the stock return difference; $Growth_{i,t}$ is the earnings growth rate; $ADV_{i,t}$ is the advertisement expense; $Mfcount_{i,t}$ is the total forecasts times by analysts; η_i, g_t, and $v_{i,t}$ are dummy variables.

3 Results and Analysis

In this section, we present the performance of our automatic model for disclosure quality assessment, and the regression analysis of Chinese report readability.

Table 2 presents the average predictive accuracy from 10-fold cross validation of our automatic assessment model, using different term-weighting schemes and classifier constructions.

Table 2. Average Accuracy of Four-Class Classification Models

	Classifier Models	
Weighting Schemes	**SVM-score**	**SVM-prob**
atn	0.61542	0.61605
ltn	0.62192	0.61793
atc	0.61751	0.61603
ltc	0.62192	0.61604

As we observe from the Table 2, the choice of different multi-class label assignments methods do not perform significantly differently from each other. Nor does the different weighting schemes. We remind our readers that this classification is based on analysts manual ratings as gold standard. When compared with other research[15,16,17], ***the best accuracy achieved at 62.19% for Chinese reports in fact is about 10% improvement over the performance of other classification research on English reports using financial indicators for class definitions***. We pick SVM-score model with ltn weighting scheme to look into the binary classification performance for each of the four quality ratings, namely "Excellent", "Good", "Pass", and "Fail". Results are shown in Table 3.

Table 3 shows ***higher classification accuracy for each class than for overall 4-class classification***. In particular, ***the prediction for the "Fail" class reports and "Excellent" achieves the highest two accuracies***. We look into the details of the prediction with a contingency table analysis of SVM-score model with ltn weight. As shown in Table 4, the true number percentage of "Excellent" and "Fail" reports is 13.45% of the total sample set. Although the predictions for the "Excellent" and "Fail" categories of annual reports achieve the highest accuracy up to 95%, the percentage of these two prediction is only 5% of the model's total predictions. This implies that the multi-class model is able to ***identify "Excellent" or "Fail" quality reports with good precision***, but inefficient in identifying the majority of the "Excellent" or "Fail" quality reports. Another observation is that the two incorrect classification errors with the largest percentage occur for predicting "Pass" reports as "Good report" (18.22%) and predicting "Good" reports as "Pass" (6.94%). This indicates that ***it is more difficult for our multi-class model to distinguish between "Good" and "Pass" reports. On the contrary, our model did not make***

Table 3. Accuracy of SVM-score Binary Classifier with ltn Weights for Predicting Each Class

SVM-score Binary Classifier Models with ltn Weight				
Folds	Excellent	Good	Pass	Fail
Fold 1	92.00%	68.21%	69.68%	96.00%
Fold 2	90.95%	67.79%	70.95%	95.79%
Fold 3	92.21%	65.47%	70.32%	95.79%
Fold 4	92.65%	61.34%	69.75%	95.80%
Fold 5	91.79%	66.11%	70.74%	95.79%
Fold 6	91.77%	70.04%	71.31%	95.99%
Fold 7	92.03%	68.34%	71.91%	96.44%
Fold 8	91.37%	66.53%	70.95%	95.79%
Fold 9	91.39%	65.97%	68.49%	95.59%
Fold 10	91.37%	66.53%	68.42%	95.79%
Average	91.75%	66.63%	70.25%	95.88%

Table 4. Contingency Table of SVM-score Multi-class Models with ltn Weights

SVM-score Multi-class Model with ltn Weight					
	True Excellent	**True Good**	**True Pass**	**True Fail**	**Total**
Predicted Excellent	2.69%	1.24%	0.27%	0.00%	4.21%
Predicted Good	6.29%	47.53%	18.22%	1.07%	73.11%
Predicted Pass	0.21%	6.94%	11.74%	2.95%	21.84%
Predicted Fail	0.00%	0.11%	0.51%	0.23%	0.84%
Total	9.20%	55.82%	30.74%	4.25%	100.00%

any mistakes in predicting "Excellent" as "Fail" (0%) or predicting "Fail" as "Excellent" (0%).

About our regression analysis on the association of Chinese report readability and analysts effort, we present our results in Figure 1. Models 1 and 4 are fixed effect models without controlled variables. Models 2 and 5 are fixed effect models with controlled variables. The significantly negative coefficient values indicate the negative association between the readability measure (which indicates the level of difficulty in reading) and the number of analysts following the reports. Models 3 and 6 are fixed effect models with controlled variables and dummy variable. The coefficient values are still negative, although the association is not significant. These results indicate that ***analysts effort in following annual reports is negatively associated with the level of difficulty in reading the reports. In other words, easier to read annual reports attract more attention from analysts in their evaluation.***

Independent Variables	Dependent Variable: *ANALYST*					
	(1)	(2)	(3)	(4)	(5)	(6)
Readability (The first index)	-.0764401***	-.0751219***	-.0022596			
	-8.86	-10.3	-0.11			
Readability (The second index)				-.0813372***	-0.0798391***	-.0024844
				-8.87	-10.29	-0.11
Logsize		0.0215916	-0.1725738		0.0214627	-0.1725269
		0.37	-1.48		0.37	-1.48
Pinst		0.02735339***	0.0598901**		0.027352***	0.059891***
		4.96	5.09		4.96	5.09
Lsegments		0.0429698	-0.462776*		0.0424126	-0.46279*
		0.16	-2.01		0.16	-2.01
Std_red		3.446439***	1.041596**		3.445953****	1.041664**
		4.85	3.90		4.85	3.9
Growth		0.0072067	-0.1158578*		.0072127	-0.1158493*
		0.09	-2.38		0.09	-2.38
ADV		2.178051	3.156798		2.186037	3.158065
		0.7	1.41		0.7	1.41
Mfcount		0.2687364***	0.2343682***		0.268734***	0.2343686***
		42.32	10.61		42.32	10.61
Prob >chi2	0	0	0	0	0	0
Within R^2	0.0419	0.3908	0.4599	0.0419	0.3908	0.4599
obs	4847	4847	4847	4847	4847	4847
Model	F	F	F,T	F	F	F,T
Methods	OLS	OLS	OLS	OLS	OLS	OLS

Fig. 1. Regression Analysis of Report Readability. Readability one is the 3-factor readability index. Readability 2 is the 7-factor readability index. Models 1 and 4 are fixed effect models without controlled variables. Models 2 and 5 are fixed effect models with controlled variables. Models 3 and 6 are fixed effect models with controlled variables and dummy variable.

4 Conclusion

We presented a series of experiments designed to explore the feasibility of constructing automatic assessment system for evaluating the information disclosure quality in annual reports. In contrast to the evaluation of English annual reports using financial performance indicators as surrogates, we exploit the manual ratings from analysts to train our learning classifiers. Our model for overall four-class classification achieves better performance to the extent of classification accuracy than the counterpart research on English reports. We speculate that the use of manual ratings could serve as better guidelines for automatic assessment of disclosure quality than financial or accounting measure.

Further analysis of the classifiers performance shows that distinguishing between "Excellent" versus "Fail" quality reports is much more efficient than between "Good" and "Pass" quality reports. Our current methods could supplement analysts manual process in identifying "Excellent" and "Fail" reports. Future research may be directed towards performance improvement of evaluating "Good" and "Pass" reports.

We further calculates the readability measure (i.e. level of reading difficulty) for Chinese annual reports. We studied the association of readability with analysts following effort. Our findings suggest that easier to read report may attract more analysts attention in following and analyzing the reports. Our study on how readability index and its component factor are related to disclosure quality is ongoing. Results will be presented in our future study.

From this study, we conclude that exploiting the manual ratings to develop automatic assessment model for disclosure quality not only is highly feasible, but also can supplement manual evaluation process. Our findings have give us a better understanding of the opportunities and challenges in automatic assessment of disclosure quality and prepare us for future work in this direction.

Acknowledgments. We thank the workshop participants at CISCO School of Informatics and the School of Finance at Guangdong University of Foreign Studies for their helpful comments and support. This work was partially supported by Grant 12YJAH103 from the Ministry of Education of China Project, and Grant 2011J5100004 from Guangzhou Science and Technology Program Project.

References

1. Wei, W., Gaofeng, J.: Information Disclosure, Transparency and the Cost of Capital. Economic Research Journal 7 (2004)
2. Ying, Z., Zhengfei, L.: The Relationship between Disclosure Quality and Cost of Equity Capital of Listed Companies in China. Economic Research Journal 2 (2006)
3. Core, J.E.: A Review of the Empirical Disclosure Literature: Discussion. Journal of Accounting and Economics 31(13), 441–456 (2001)
4. Davis, A., Piger, J., Seor, L.: Beyond the Numbers: An analysis of optimistic and pessimistic language in earnings releases. Working paper, Washington University in St. Louis (2006)
5. Li, F.: Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45(2-3), 221–247 (2008)
6. Li, F.: The information content of forward-looking statements in corporate filings A nave Bayesian machine learning approach. Journal of Accounting Research 48(5), 1049–1102 (2010)
7. Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A.: Predicting risk from financial reports with regression. In: NAACL 2009 Proceedings of Human Language Technologies, pp. 272–280 (2009)
8. Feldman, R., Govindaraj, S., Livnat, J., Segal, B.: Managements tone change, post earnings announcement drift and accruals. Review of Accounting Studies 15(4), 915–953 (2010)

9. Lehavy, R., Li, F., Merkley, K.: The effect of annual report readability on analyst following and the properties of their earnings forecasts. The Accounting Review 86(3), 1087–1115 (2011)
10. Gelb, D.S., Zarowin, P.: Corporate disclosure policy and the informativeness of stock prices. Review of Accounting Studies 7, 33–52 (2002)
11. Wang, X., Shen, W.: The Empirical Research on the Relationship between Control Structure and the Quality of Information Disclosure. Securities Market Herald 4 (2008)
12. Chen, Q.: Disclosure Quality and Market Liquidity. South China Journal of Economics 10 (2007)
13. Yang, S.-J.: A Readability Formula For Chinese Langauge. Ph.D. thesis. The University of Wisconsin (1971)
14. Platt, J.C.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press, Cambridge (1999)
15. Balakrishnan, R., Qiu, X.Y., Srinivasan, P.: On the predictive ability of narrative disclosures in annual reports. European Journal of Operational Research 202, 789–801 (2010)
16. Qiu, X.Y.: Towards building ranking models with annual reports. Journal of Digital Information Management 8(5), 338–343 (2010)
17. Qiu, X.Y., Srinivasan, P., Hu, Y.: Supervised Learning Models to Predict Firm Performance with Annual Reports. Journal of the American Society for Information Science and Technology (forthcoming)